图书情报工作
第63卷 第16期 2019年8月

研究数据使用统计新标准及其应用案例研究*

■ 林伟明　叶兰
深圳大学图书馆　深圳 518060

摘要：［目的/意义］介绍 Make Data Count 与 COUNTER 联合推出的《研究数据使用统计实施规范》，为数据级别计量提供新指标与新视角。［方法/过程］通过对标准文本的分析，介绍该规范的提出背景、目标、范围、相关概念及核心内容，通过案例剖析 Dash、DataONE、Zenodo 及其他 7 个数据存储库对《规范》的应用情况。［结果/结论］研究数据的使用统计具有其独特之处，《规范》的推出可对数据引用及数据替代计量形成补充进而描述完整的科研学术影响力。目前遵循该规范的数据存储库还不多，为推动数据使用计量的应用，需要标准组织、科研人员、机构库及数据存储库、出版商、科研机构及资助机构、图书馆等不同利益相关者在数据产生、管理、传播与利用等环节的相互合作。

关键词：研究数据　使用指标　使用统计　使用数据　数据级别计量
分类号：G251.4
DOI：10.13266/j.issn.0252-3116.2019.16.004

随着数据密集型科研范式的发展，研究数据越来越成为学术产出的重要组成部分。为鼓励科研人员公开、共享与重用研究数据，学术界已开始呼吁像对待研究论文一样对待研究数据，将研究数据纳入科研评价的对象，探索研究数据的计量与影响力评价。研究数据的使用统计被科研人员及其他利益相关者认为是重要的指标之一，仅次于数据引用次数[1]。但是，由于缺乏相应的标准规范研究数据的使用统计数据的收集与获取，研究数据的使用统计指标还未发挥其应有的作用。为此，Make Data Count 项目与 COUNTER（Counting Online Usage of Networked Electronic Resources）项目合作开发，于 2018 年 6 月 5 日正式发布《研究数据使用统计实施规范》（第 1 版）[2]（以下简称《规范》），规范研究数据的使用统计的生成与发布，为数据存储库及数据平台提供者提供一致性、可靠性和相互兼容性的关于研究数据的使用统计，促进数据存储库、图书馆、基金资助者及其他利益相关者推动研究数据的重用。本文介绍新版研究数据的使用统计实施规范的提出背景、基本内容及其应用情况，并基于利益相关者提出推动数据使用计量的建议。

1　数据使用统计提出背景

根据美国国家信息标准化组织 NISO 的定义[3]，数据使用是用户访问以及下载一个公开出版的数据集的行为，其统计范围包括数据的下载、数据访问、数据集标注等。数据使用统计的提出是利益相关者意识到其重要性，并考虑建立综合的数据影响力评价的需求，以及为弥补数据使用指标领域现实空缺的情况下提出的。

1.1　数据使用指标具有重要性

数据使用指标可以帮助科研人员在数据正式被引用发生前就了解其研究数据的受关注程度，并作为重要的科研评价数据，激励科研人员参与数据共享与重用；帮助机构监测数据的使用趋势进而评估数据存储设施的服务效果，同时评估存储设施与网络系统的容量需求，还可针对受欢迎度高的数据集创建专门的馆藏；帮助数据存储库了解其数据的使用情况及某些特定数据集的影响力；帮助基金资助机构了解其所资助的科研产出（研究数据）对科学进展及整个社会的贡献。

1.2　建立综合的数据影响力评价的迫切需求

尽管数据引用指标是当前数据计量及数据影响

* 本文系教育部人文社会科学研究青年基金项目“基于成熟度视角的高校图书馆科学数据管理服务能力评价研究”（项目编号：19YJC870028）研究成果之一。
作者简介：林伟明（ORCID：0000-0003-4287-1456），馆员，硕士；叶兰（ORCID：0000-0002-3079-5399），副研究馆员，硕士，通讯作者，E-mail：yel@szu.edu.cn。
收稿日期：2019-01-08　修回日期：2019-04-09　本文起止页码：32-42　本文责任编辑：杜杏叶

力评估领域首当推崇的测量指标，但其并不能全面说明数据重用的概貌。科学文献领域已开始反思引用作为学术评价计量指标的单一性和绝对性，进而提出了基于引用、使用及替代计量等多种方式来综合评价学术影响力。对于研究数据，国外已有研究提出建立多种数据集评价指标的建议，如 K. M. Fear[4]指出数据集的评价计量不能依据单一指标，应多因素考虑，如数据引用计量、二次影响(如 G 指数)、数据重用的学科广度以及数据下载量。J. Bollen 等[5]提出基于使用且覆盖整个研究过程的影响计量方法，计量内容涵盖引用、发现、下载、同行评议邮件数、阅读以及保存等。因此，数据使用指标的推出可与数据引用及替代计量指标形成补充进而描述完整的数据影响力。

1.3 现实中缺乏数据使用指标的最佳实践

数据引用、数据使用(下载及浏览)、数据的替代计量是数据计量与影响力评价的三大指标体系，其中，数据引用计量是开展最早且研究最多的领域，其次是数据的替代计量。目前，数据引用、数据的替代计量已经公布了一定的规范与标准。如未来科研交流与电子学术组织(The future of research communication and e-Scholarship，FORCE11)在 2014 年推出了《数据引用原则》(Joint declaration of data citation principles)[6]。Scholix 项目(A framework for scholarly link eXchange)[7]通过建立框架推动学术文献与数据之间链接信息的交换，帮助科研人员了解学术文献中的数据及引用数据的学术文献。科睿唯安已于 2012 年推出了数据引文索引(Data citation index)跟踪和记录单个数据集的引用次数。美国国家信息标准化组织 NISO 于 2013 年开展替代计量项目，其中的一个工作组主要研究数据集、软件等非传统科研成果的替代计量[8]。目前，数据使用指标的最佳实践仍在探索之中。为推动数据使用指标的应用，Make Data Count 项目从 2014 年起在美国国家科学基金会(National science foundation，NSF) EAGER 基金资助下，基于 2009 年由 PLOS 启动的“开放资源文章级计量指标项目”(Article-level metrics，ALM)-Lagotto 开展数据级计量指标 (The data-level metrics，DLM) 试点项目[9]。最初成员包括 PLOS、加州数字图书馆(CDL)及 DataONE，PLOS 在后来退出，DataCite 加入，并最终推出《研究数据使用统计实施规范》，使各不同的数据存储库之间按照统一的规范提供使用数据，这是实现利用研究数据使用统计了解研究数据如何重用这个过程中的重要里程碑，对目前数据引用领域中的最佳实践与服务也是一个有效的补充。

2 数据使用统计研究现状

国内外研究目前主要从整体角度研究数据级别的计量与数据影响力，较多关注科学数据影响力的引文指标，单独从使用统计角度研究数据级别的计量还不多。具体如下：

2.1 从整体角度研究数据级别的计量与数据影响力

K. M. Fear[10]提出 5 个用于评估科学数据影响力的指标，包括数据重用频次、重用数据的出版物的质量、重用数据的出版物的多样性、源于单个数据集的相关网络规模以及数据集的下载数量。此外，部分国外组织及实践项目早已开展相关研究。如 Knowledge Exchange 在 2013 年发布的《研究数据的价值》报告从文化与技术角度分析数据计量的概念、与数据计量相关的数据共享、数据共享和数据计量的利益相关者、相关的知识库与工具等问题[11]。英国数据监护中心(The digital curation centre，DCC)在 2015 年发布《如何利用计量指标追踪研究数据的影响力》报告提及数据计量的相关概念、相关工具与服务、数据计量面临的挑战等[12]。科研管理信息标准推进委员会(The consortia advancing standards in research administration information，CASRAI)成立数据集级别计量课题小组(Dataset level metrics subject group)旨在集结不同利益相关者共同研制数据级别的计量指标[13]。研究数据联盟(The research data alliance，RDA)和世界数据系统(The world data system，WDS)联合成立了数据出版计量工作组(RDA/WDS publishing data bibliometrics WG)研究数据计量指标及相应服务。以上实践项目基本在 2013-2015 年之间开展，目前已极少更新。它们推动并引发科学界对数据级别计量的思考，但由于数据计量涉及的问题较为复杂，尚未形成系统完善的评价方法。国内主要侧重于介绍国外项目与进展，较早研究数据级别计量的是顾立平[14]，介绍数据级别计量的概念、发展与应用。王毅萍[15]介绍科学数据影响力的内涵、类型、关系、相关主体及评价方法。孟阳[16]分析对比数据计量与文献计量之间的异同。

2.2 从数据引用计量角度研究数据级别计量

国外无论是在理论还是实践上都对数据引用开展了深入研究，不仅有 DataCite、研究数据联盟、英国数据监护中心等组织建立数据引用标准与引用原则，还推出了数据引文索引工具跟踪和记录引用次数。我国也

推出了《科学数据引用》国家标准，另有不少研究关注国外的数据引用规范，同时利用数据引文索引工具分析社会科学数据的影响力，如丁楠[17]、邢红梅[18]等。

2.3 从数据使用角度研究数据级别计量

学术界较早正式提出数据使用统计并将其作为单独对象进行研究的是 P. Ingwersen 和 V. Chavan[19] 在 2009 年提出的数据使用索引（Data usage index），并以 GBIF（生物多样性数据库）的数据为基础，构建了包含搜索密度、下载密度、使用影响、兴趣影响等 14 个指标在内的数据使用指标。遗憾的是这套指标仅适用于 GBIF 数据存储库，其科学性、普适性有待进一步研究。国内目前仅丁培[20]涉及科学数据使用统计，研究科学数据使用统计的收集、规范、清洗、报告等关键流程。

综上所述，鉴于目前国内外的研究较多关注数据计量的引文指标，较少涉及数据的下载、浏览等使用指标。本文推介 Make Data Count 项目与 COUNTER 为制定普适性的研究数据使用统计指标而推出的《研究数据的使用统计实施规范》，以进一步发挥数据使用统计指标在数据计量中的作用。

3 《规范》简介

《规范》是由在数据管理领域具有丰富经验的三个机构与学术资源使用统计权威组织 COUNTER 共同研制的成果。COUNTER 项目是 2002 年 3 月启动的一项国际首创计划，目的是规范数据库商向图书馆提供的使用数据格式、内容、术语等，使各数据库商生成的使用数据具有一致性、可靠性和相互兼容性，并且方便记录和交换[21]，其主要针对电子期刊、电子图书、数据库、多媒体等学术资源的使用统计。Make Data Count 是由斯隆基金资助的为期两年的项目，由加州数字图书馆（California digital library）、DataCite 及 DataONE 组成。其中，加州数字图书馆是由加州大学在 1997 年成立，数据监护中心（University of California curation center，UC3）是 CDL 的四大主要项目之一，帮助研究人员及加州大学图书馆对数字资产进行管理、保存与访问，并提供数据生命周期的管理工具与服务。DataCite 成立于 2009 年底，是为研究数据提供永久标识符 DOI 的国际性非营利组织，帮助研究社区定位、识别及引用研究数据。DataONE（Data observation network for earth）于 2009 年 8 月启动，是 NSF 资助的 DataNet 项目之一，为描述与发现地球观测数据建立一个分布式框架及可持续的网络基础设施。Make Data Count 和 COUNTER 成员自 2017 年 6 月开始讨论研究数据使用统计的推荐标准，并于一年后推出该规范。

3.1 数据集相关概念

《规范》统计的是数据集的使用情况。以下 4 个概念是《规范》对数据集及其上位类、下位类所规范的定义，有助于了解研究数据的结构及细粒度，确定《规范》所统计的对象。

数据集（dataset）是由某个代理商出版或保管的数据的集合，与其元数据一起，按一种或多种格式提供访问或下载[22]。数据集是 COUNTER 中的一个内容项。与其同义的词是数据包（data package）。

数据组成部分（Component）是一个数据集中的某个数据，可单独提供访问或下载。与其同义的词是数据文件（data file）、数据颗粒（data granule）。

数据集集合（Collection）是数据集的集合。相关的术语是目录（catalog）、存储库（repository）。

数据集的版本（Version）是数据集的基本特征，是指一个数据集的多个版本。内容或（与）元数据的变化、一个或多个组成部分的变化以及可能导致组成部分固定属性的变化都会产生不同的版本。

3.2 《规范》参考的标准

《规范》是在参照电子资源使用统计、使用统计收割、引用、替代计量等多个已有标准的基础上提出的。首先参照了 2017 年 7 月正式发布的《第 5 版 COUNTER 实施规范》[23]（COUNTER code of practice release 5）。COUNTER 是主要针对期刊、图书等学术资源的使用统计标准，因而其中很多定义、处理规则及报告建议都可适用于研究数据。此外，参考了电子资源使用统计收割标准 SUSHI（ANSI/NISO Z39.93－2014：Standardized usage statistics harvesting initiative）[24]。该标准代替人工来收集使用数据的统计报告，同样适用于研究数据。另外，还参考了“Scholix metadata schema for the exchange of scholarly communication links”[25]以规范描述数据集的元数据，及美国国家信息标准化组织的《替代计量项目成果》（NISO RP－25－2016：outputs of the NISO alternative assessment metrics project）[3]对数据计量以及永久标识符的相关推荐。

3.3 《规范》的目标、范围、与《第 5 版 COUNTER 实施规范》的关系及其管理

《规范》目的是为数据存储库及数据平台提供者提供一致性、可靠性和相互兼容性的关于研究数据的使用统计。

目前，《规范》涉及的对象只是数据集层面的使用统计，未来将根据用户需求与反馈提供数据集中所有

组成部分的使用统计。《规范》主要对统计的数据元素、数据元素的定义、使用报告的内容与格式、数据处理要求、避免重复计量等内容进行规定。

《规范》由研究数据管理领域人员与 COUNTER 合作完成，并遵循《第 5 版 COUNTER 实施规范》，仅在必要时与《第 5 版 COUNTER 实施规范》有所不同。如研究数据不需要提供机构层面的使用统计，但是倾向于按地理位置划分使用数据。另一个显著不同是其版本，需要整合某个数据集所有版本的使用统计。此外，也不需要按照文件格式发布统计报告，如不单独提供 CSV 或 XLSX 格式的下载量。

《规范》由 Make Data Count 项目与 COUNTER 项目合作开发，也由其合作管理。

3.4 《规范》的核心内容

《规范》共包括 8 个部分内容：①前言；②总览；③报告的技术实施；④使用报告；⑤报告的传递；⑥使用数据收集方式；⑦数据处理；⑧利用 SUSHI 自动收割报告。其中，第 3、4、5、7、8 是该实施规范在执行过程中的核心内容。

3.4.1 第 3 部分 报告的技术实施 该部分介绍必须提供的报告，描述所有报告的通用格式，定义报告属性及其赋值。

关于提供的报告，报告名称为“Dataset Master Report”，是对数据集层级的使用行为的细粒度与个性化报告，允许用户应用过滤器及选择各种配置选项。这个报告适用于存储库（repository）及数据存储库（Data repository）。存储库和数据存储库是《规范》规定的两种托管类型（Host type）。存储库是托管包括研究数据在内的多种研究产出类型的仓储，如机构库即属于这种类型，如 Figshare。数据存储库是仅托管研究数据的存储库，学科领域的数据存储库属于这种类型，如 CDL Dash、Dryad 等。

报告的格式可以是表格格式或机器可读的 JSON（JavaScript object notation）文件格式。

所有报告的结构基本相同，都包含一个表头，与《第 5 版 COUNTER 实施规范》不同的是，研究数据使用统计报告没有机构相关的元素，即机构名称（Institution_name）及 ID（Institution_ID）。表头元素包括 10 个：①报告名称（Report_name）；②报告 ID（Report_ID）；③版本；④指标类型（Metric_types）；⑤报告过滤器（Report_filters）；⑥报告属性（Report_attributes）；⑦例外（Exceptions）；⑧报告日期（Reporting_period）；⑨报告创建时间（Created）；⑩报告创建者（Created_by）。

其中，统计指标类型是最重要的报告元素。《规范》参照《第 5 版 COUNTER 实施规范》使用“调查量（Investigations）”及“请求量（Requests）”指标。这两个统计指标是第 5 版 COUNTER 新引入的指标。“调查量”表示一个用户访问某内容项的信息（如一篇文章的文摘或详细的描述性元数据）或某内容项本身（如一篇文章的全文）。“请求量”是指用户请求某内容项的全文的次数，通常以浏览、下载、email 或打印等形式呈现。为清楚区分调查量与请求量，《第 5 版 COUNTER 实施规范》提供了一个调查量与请求量的关系图（见图 1）。从中可看出，请求量是调查量的一个子集。应用于研究数据，可理解为：任何适用于数据集的用户行为（包括元数据）都可认为是“调查量”，包括某个数据集的下载或浏览量。而“请求量”仅表示检索或浏览数据集本身的用户行为。“调查量”与“请求量”又各分为“Total”与“Unique”（具体见表 1）。

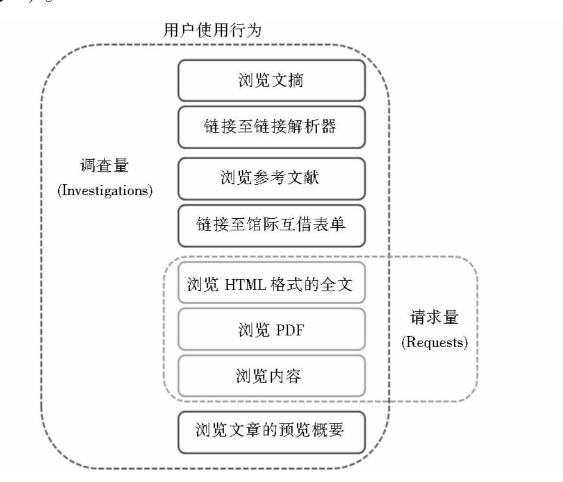


图 1 调查量与请求量的关系

3.4.2 第 4 部分 报告 该部分提供每种报告的详细规范及报告所包含的元素。《规范》目前只提供一种报告，即数据集报告。标准的数据集报告应包含以上 10 个表头元素，且表头元素应严格按照顺序出现（参照上一段描述中出现的顺序），表头标签的拼写（如大小写）都有严格要求，不可随意改动。除了表头元素外，如果该数据集有的话，必须包含数据集名称、出版商、创建者、出版日期、数据集版本、DOI、URL 等信息。其中，在表格格式的报告中必须包含 DOI 或其他 ID 信息或 URL。

3.4.3 第 5 部分 报告的传递 该部分说明内容提供者必须提供的信息以确保其报告能被用户获取。《规范》对使用报告的传递作出如下规定：第一，报告必须

表 1 《研究数据的使用统计实施规范》(第 1 版)的统计指标类型

统计指标类型	详细描述	适用的场景	适用的报告
总数据集调查量(Total_dataset_investigations)	某个数据集被访问的全部次数及传输的数据量(以兆字节统计)。 * 重复点击过滤器适用于该统计指标。 * 提供每个版本的数据集的调查量(包括次数及数据量),并提供所有版本的总量。	存储库 数据存储库	DSR(Dataset master report)
数据集调查量(Unique_dataset_investigations)	在某个特定的用户登陆时段内(通常指 1 小时时间窗口)数据集的调查量。在同一用户登陆时段,对同一个数据集的多个组成部分的访问,仅统计一次对数据集的调查量。 * 提供每个版本的数据集的调查量(包括次数及数据量),并提供所有版本的总量。	存储库 数据存储库	DSR
总数据集请求量(Total_dataset_requests)	某个数据集被检索(指内容的全文或组成部分被访问或下载)的全部次数及传输的数据量(以兆字节统计)。 * 重复点击过滤器适用于该统计指标。 * 提供每个版本的数据集的请求量(包括次数及数据量),并提供所有版本的总量。	存储库 数据存储库	DSR
数据集请求量(Unique_dataset_requests)	在某个特定的用户登陆时段内(通常指 1 小时时间窗口)数据集的请求量(包括次数及数据量)。在同一用户登陆时段,对同一个数据集的多个组成部分的访问,仅统计一次对数据集的请求量。 * 提供每个版本的数据集的请求量(包括次数及数据量),并提供所有版本的总量。	存储库 数据存储库	DSR

是以下两种格式:① TSV(Tab separated value)格式文件,可容易且无误差或无数据遗漏地导入至 EXCEL 表格中;②JSON 格式,且遵循"研究数据 SUSHI API 规范"(Research data SUSHI API specification);第二,报告必须能以单个文件传递以方便报告的自动处理;第三,表格版本的使用报告应能上载到一个以密码控制的网站上,用户使用密码能够随时访问,当使用数据有更新时,能够通过邮件提醒用户;且能提供过滤器或选项供用户选择,并必须提供平台的所有标准报告的浏览;第四,至少每月提供一次使用报告,使用报告应在上个月报告发布后的 1 个月之内更新,使用数据通常按整月来处理,但如果不足整月,也可输出部分使用数据;第五,至少要保留本年度以及之前两年内的使用数据;第六,报告必须能够通过电子资源使用统计收割标准 SUSHI 协议收割。

3.4.4 第 7 部分 底层数据的处理原则 该部分说明了使用统计中数据采集和处理原则,主要讨论统计数据的返回码、重复点击的过滤、机器人及爬虫检索等相关的问题。

关于重复点击过滤问题,规定同一用户在一个链接上间隔不足 30 秒的双击只被记为一次点击。如第一次点击发生在 10.01.00,第二次点击发生在 10.01.29,这被认为是重复点击,只记录一次点击。如果第一次点击发生在 10.01.00,第二次点击发生在 10.01.35,这被认为是两次单独的点击,记录为两次点击。重复点击可通过鼠标点击或按更新或返回按钮触发。当在一个 URL 上发生间隔不足 30 秒的两次行为,第一次请求必须清除,而保留第二次的请求信息。对于如何判断是否是同一用户的点击,《规范》提供了 4 种方式(按照可信度从高至低排列):①根据用户登陆时的用户信息判定,如用户名;②根据用户 cookie 来辨别;③根据 session cookie 来判断;④通过 IP 及浏览器的用户代理来判断。

关于机器及网络爬虫检索等相关的问题,《第 5 版 COUNTER 实施规范》强调真正的人类用户的使用量,过滤了已知的网络爬虫、网络蜘蛛等的使用量,同时允许通过脚本语言(如 python、curl、wget 及 Java)或自动工具等合法的机器检索,这同样适用于研究数据中。《规范》允许合法的机器检索,体现在"检索方法"(Access_method)这个报告属性中,通过赋值"Regular"或"Machine"来区别合法的机器浏览或下载量,但是不允许也不统计通过网络爬虫或网络机器人的检索量,并通过黑名单来排除用户通常使用的爬虫或机器人代理,可参见第 5 版 COUNTER 所列出的网络机器人或爬虫列表。

3.4.5 第 8 部分 利用 SUSHI 自动收割报告 该部分提供对 SUSHI 支持的详细描述。电子资源使用统计收割标准 SUSHI 是为图书馆更为高效地收集符合 COUNTER 标准的使用数据而推出的数据采集与传输标准协议,解决和实现了图书馆电子资源使用数据的自动化收割和管理问题。从 2008 年第三版 COUNTER 发布起即将 SUSHI 纳入 COUNTER 标准中,并作为遵循 COUNTER 标准的必要条件之一。参照第 5 版 COUNTER,《规范》同样要求内容提供商必须支持其报告可通过 SUSHI 自动收割,并制定了"研究数据 SUSHI API 规范"。

4 研究数据使用统计的独特之处

通过梳理《规范》的内容,笔者发现部分使用统计数据的处理与输出方式是研究数据所独有的。《规

范》也特别强调了其独特之处，主要体现在：

首先，不按照机构划分使用量。因为研究数据不像电子期刊、电子图书等学术资源是通过订阅购买方式获得，在研究数据领域，订阅购买方式的发生不是很普遍，因此，按照机构区分使用量的意义不大。为满足对研究数据使用统计的地理分布信息的需求，通过国别而不是机构来提供使用报告，这比按照机构划分使用量更利于使用数据的公开与共享。

第二，按照地理信息（国别）而不是 IP 地址提供报告。在研究数据领域，按照国别地理信息划分使用统计比按照 IP 来划分更有意义，因为提供地理信息可帮助了解同一数据集在不同地理位置的使用情况。对于数据集来说，其使用取决于某个地理位置的用户，如描述特定地区的数据集。对于美国等大国，使用报告可以提供州或省级别的统计数据。

第三，提供各版本的使用统计。与其他学术资源相比，版本在研究数据中应用普遍且较为复杂。《规范》特别强调了版本，建议为每一个特定版本的研究数据输出相应的使用报告，并统计所有版本的总使用量。

第四，不提供各种格式的使用量，而是提供数据量（data volume）。与基于文本的学术资源相比，研究数据可从多种类型的文件格式中检索。《规范》没有按照文件格式划分对研究数据的请求量，如不单独提供 CSV 或 XLSX 格式的下载量，而把请求的数据量作为使用报告的一部分，主要是考虑这个变量在研究数据中比在其他学术资源中意义更大。请求的数据量大小配合请求及调查量可有助于对比数据存储库之间在数据打包方面的差异，便于比较不同细粒度的数据集的使用统计。

5 研究数据使用统计的应用案例分析

《规范》主要应用于机构库及数据存储库中。机构库及数据存储库应用该标准规范需要完成 5 个步骤[26]：①阅读并了解《规范》；②按照该标准规范处理使用日志；③发送处理好的标准化使用日志至一个开放的中心（目前是 DataCite Hub 作为研究数据使用统计的开放中心）；④从该开放中心提取使用及引用指标；⑤在存储库平台展示标准化的使用及引用指标信息。

5.1 加州数字图书馆 Dash 和 DataONE

自《规范》制定以来，项目团队的两个存储库——Dash（CDL）和 DataONE 实施了标准化的数据使用和引用指标。其中，数据使用指标主要包括浏览及下载量，根据《规范》处理内部日志，并将标准格式化的使用日志发送到 DataCite Hub 以供公共使用并最终进行整合。引用指标信息则来自 CrossRef Event Data。

加州数字图书馆 Dash 是为研究人员提供描述、上传、管理与分享其研究数据的数据存储库，其按照 DataCite 元数据方案（DataCite metadata schema）描述所有数据集，并为每个数据集提供 DOI。Dash 根据《规范》收集数据集的使用日期与时间、请求的 IP 地址、登录时段缓存 ID、用户缓存 ID、用户名或用户 ID、被请求的 URL、数据集的 DOI、数据量大小（仅用于 request 指标，investigation 指标不用统计数据量大小）等项目。同时，按照《规范》要求在提供数据集的使用统计时提供关于数据集的描述性元数据，包括题名、出版商、出版商 ID（如 ISNI 或 GRID）、创建者、出版日期、数据集版本、数据集的其他 ID（如有则可提供）、URL（DataCite 可解析的 URL）、出版物年份等。其中前三项为必备字段，其余为可选字段。在对具体统计数据的处理上，Dash 根据《规范》第 7 部分“底层数据的处理原则”，区分机器与人工两种类型的用户使用，并根据 IP 按国家级别划分使用数据[27]。

图 2 为加州大学数字图书馆项目的 Dash 页面，提供数据使用指标（包括浏览量及下载量）和数据引用指标，其中的浏览量（Views）相当于《规范》中的“investigations”指标，下载量（Downloads）相当于《规范》中的“requests”指标。此外还提供数据量的大小、数据版本、相关的数据出版物、标准的数据引用格式等信息。

DataONE 从 2018 年 7 月开始提供使用与引用指标用户界面，提供每个数据集的引用次数、下载次数及浏览次数[28]。图 3 为 DataONE 的使用与引用指标用户界面。

5.2 Zenodo

除项目团队的两个存储库外，笔者还发现 Zenodo 提供使用统计数据，其提供的数据使用指标比前两个数据存储库更为全面。Zenodo 是由欧洲核研究组织（CERN）于 2013 年 5 月成立并管理，由欧盟通过欧洲研究开放获取基础设施（Open access infrastructure for research in Europe, OpenAIRE）项目为其提供资助，旨在支持欧洲开放获取及开放数据运动。为贯彻欧盟开放数据政策，Zenodo 用于存储欧盟资助项目的研究成果，包括期刊文章、数据集、图片、软件、演示文稿等。Zenodo 从 2018 年 7 月开始提供使用统计数据[29]。

针对数据集，Zenodo 根据 COUNTER 及《研究数据的使用统计实施规范》追踪浏览量及下载量两类使用行为，每类使用行为都追踪访问者、访问者类型（人工、机器、爬虫）、国别、参考域名等信息，每三小时更新一次使用统计数据。目前，Zenodo 不仅提供浏览量及下



图 2　University of California CDL Dash 显示使用和引用指标(检索时间:2018－12－24)

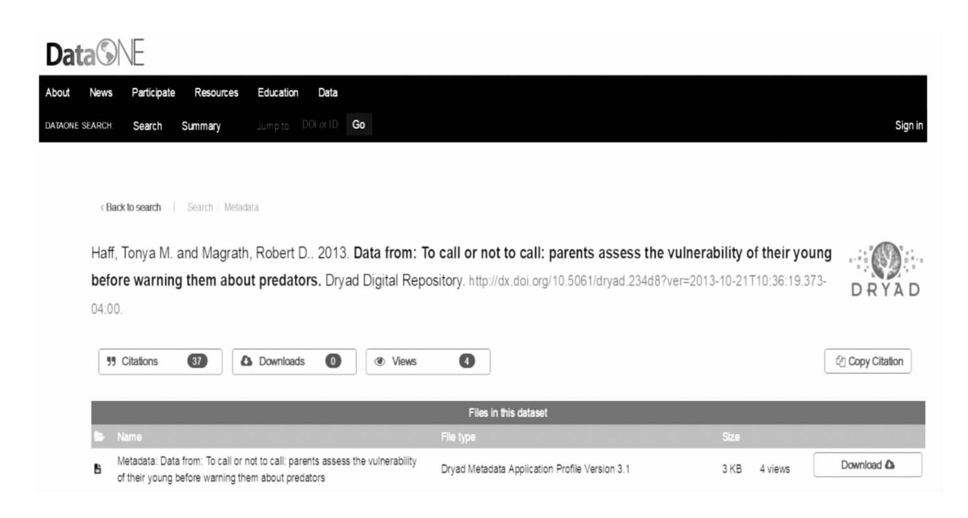


图 3　DataONE 的使用与引用指标用户界面(检索时间:2018－12－24)

载量,还提供 unique view、unique download 及数据量,并统计每个数据集版本的浏览量、下载量及数据量。这些指标的应用基本遵循《规范》中的定义。图 4 为 Zenodo 的使用指标用户界面。不过,遗憾的是其未提供引用指标。

5.3　主要大型数据存储库

为调查除 Dash、DataONE 及 Zenodo 之外其他数据存储库对使用统计指标的应用情况,笔者选取 7 个知名的大型数据存储库,包括人文社会科学领域的英国国家数据仓储(The UK data service)、美国高校校际政治与社会研究联盟(Inter university consortium for political and social research, ICPSR),自然科学领域的 GenBank(生物学)和 PANGAEA(地球与环境科学),跨学科领域的 Figshare、澳大利亚国家数据服务中心(The Australian national data service, ANDS)的 Research Data Australia 以及国际数据知识库 Dryad。通过访问以上数据存储库的网站(访问时间:2019.1.2－2019.1.5),并检索其数据集以查看是否提供数据集层面的计量指标。调查发现(见表 2),ICPSR、Figshare、Research Data Australia、Dryad 提供使用指标,其余数据存储库暂未提供使用指标,且很少有数据存储库提供引用指标及替代计量指标,仅 Figshare 提供引用指标。即使是已提供使用指标的 4 个数据存储库,除 Figshare 外,其所采纳的计量指标都还不是标准化的使用统计指标,如 Research Data Australia 和 Dryad 提供的指标并不是《规范》的统计指标类型,ICPSR 提供"Total Downloads""Total Sessions""Total Users"三个下载统计指标,每个指标都包含某一特定时段内的统计量(Unique)及重复的统计量(repeated),并提供按国别划分下载量。这与标准化的使用数据统计所涉及的指标较为接近,但仍不能算是标准的使用统计。因此,未来以上主要大型数据存储库还有待加强标准的应用。

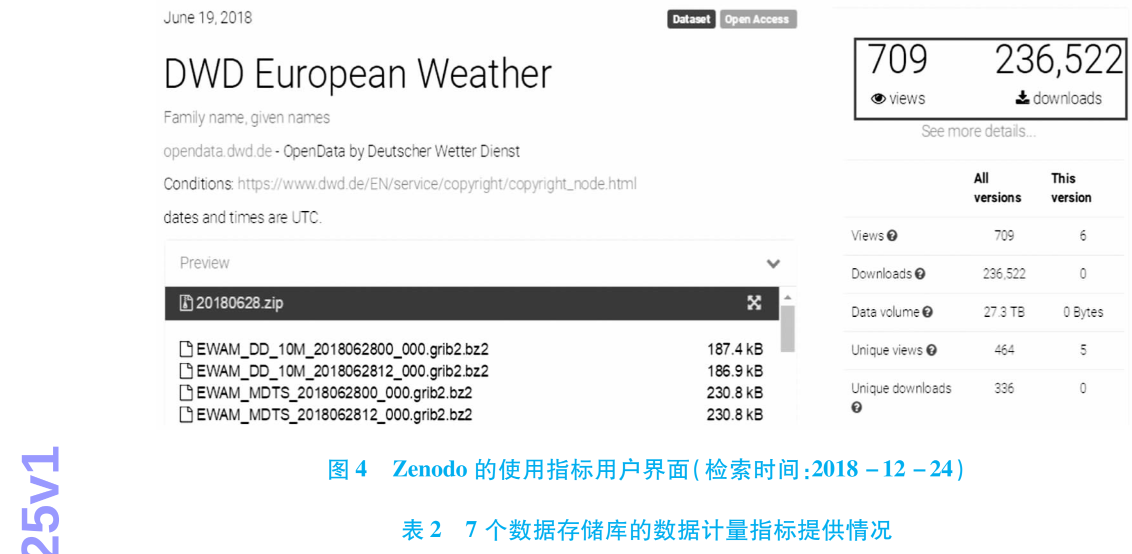


图 4　Zenodo 的使用指标用户界面(检索时间:2018－12－24)

表 2　7 个数据存储库的数据计量指标提供情况

数据存储库	使用指标	引用指标	替代计量指标
UK Data Service	无	无	无
ICPSR	下载次数	与数据集相关的出版物(并不是真正意义上的引用信息)	无
GenBank	无	无	无
PANGAEA	无	无	无
Figshare	浏览(views)、下载次数(downloads)	有	无
Research Data Australia	页面浏览(Pageviews)及 数据检索次数	无	无
Dryad	页面浏览(Pageviews)及 下载次数	无	无

6　基于利益相关者推动数据使用计量的措施与建议

数据使用是数据计量的一个分支,数据使用计量面临的最重要问题和挑战与数据计量的一般性问题紧密联系在一起。因此,本部分主要从数据计量这个更广泛的角度来推动各利益相关者采取措施促进数据计量的同时带动数据使用计量的发展。数据使用计量需要数据提供、收集、管理和传播等链条上标准组织、科研人员、机构库及数据存储库、出版商、科研机构及资助机构、图书馆等不同利益相关者在基于数据计量、开放共享、开放获取上的合作。

6.1　标准组织角度

标准组织在标准应用过程中起着重要推动作用。因此,Make Data Count 项目成员和 COUNTER 组织未来需要做好以下工作:①征集数据存储库。如前述所调研情况,目前除了项目团队的两个存储库——Dash(CDL)和 DataONE 实施了标准化的数据使用和引用指标外,很少有数据存储库提供标准化的使用统计数据。为有效推进标准的应用,需要尽可能多的数据存储库参与进来,遵循与利用该标准规范。在初期,可先吸纳一些知名的大型数据存储库(如表 2 中所提及的)加入到 Make Data Count 中,特别是已经提供使用统计的 ICPSR、Figshare、Research Data Australia、Dryad 等,鼓励其按照标准提供使用统计,之后逐步吸纳其他数据存储库。标准获得规模效应后,自然成为行业遵循的首选。②加强对《规范》的宣传与推广。可建设专门网站,提供研究数据使用统计相关的文章及新闻报道,并利用各种会议宣传数据计量。③推出标准的应用指南。研究数据的使用统计对很多存储库来说是一项新业务,需要具体指南指导其实施。目前,加州数字图书馆已制作了一个如何在存储库中应用研究数据使用统计标准规范的指南,并提供了技巧及工具[30]。此外,还专门召开了一个网络研讨会"如何在您的存储库中实施 Make Data Count"指导研究数据使用统计的应用。这对于推动标准应用具有重要作用。④推动各标准的配合使用。标准的使用并不是孤立的,《研究数据的使用统计实施规范》需要数据引用、替代计量等标准配合使用,从而建立基于使用、引用、替代计量等指标在内的综合的数据影响力评价体系。⑤保证标准持续维

护。标准通常每隔四至五年需要根据现实需求进行升级与更新。《研究数据的使用统计实施规范》目前是第一版,仍有许多需要完善的地方,如还未明确要求对内容提供商的审核。因为目前还不知道未来哪些机构愿意接受审核,对于审核程序是否完全参照《第 5 版 COUNTER 实施规范》或者有所不同还未考虑清楚。因此,未来的版本需要对内容提供商的审核细节进行规定。

6.2 科研人员角度

科研人员是数据使用计量最直接的利益相关者。如果没有科研人员共享研究数据,数据计量将是无米之炊,因此,科研人员应如同对待文章一样重视数据计量指标,积极参与数据共享。数据共享可以多种方式开展,一是可将数据存储至可信、持久、可持续的数据存储库中。使用、引用、替代计量指标的追踪依赖于数据集拥有一个稳定的存储地点,而获得稳定存储的有效方式即是将数据存储至能够实现长期保存并提供未来重用的数据存储库中。科研人员在选择可信任的合适的数据存储库时可参考 DCC 提供的标准[31]:①该存储库是否有声誉?是否经过认证?是否是研究人员所在期刊、机构或资助机构所推荐或要求的?②该存储库是否在同行中使用普遍?是否能有利于数据集的发现与重用?③该存储库是否对数据的质量进行评估?④该存储库是否为数据集提供永久标识符?⑤该存储库是否收集使用统计?⑥该存储库是否被 Data Citation Index 或其他数据库索引?二是正式出版,将数据作为正式科研成果,进行同行审议,公开出版,供其他人共享,如数据论文出版、附录数据等。《科学数据价值》报告认为数据期刊(即数据论文)是最适用于提取及发展科学数据影响力计量体系的出版方式,因为数据论文出版类似于传统科学出版物出版,可以充分运用现存的使用、引用和替代计量指标。因此,从影响力角度来看,鼓励科研人员以数据论文形式在数据期刊上发表数据。

6.3 机构库及数据存储库角度

机构库及数据存储库是《研究数据的使用统计实施规范》的主要实施者。对机构库及数据存储库而言,准确详细的描述数据集是其首要任务,因为数据首先必须能够被发现、被理解、被重用才能体现出其影响力。因此,机构库及数据存储库首先应提供有助于数据集发现的元数据,如最基本的发现型元数据包括题名、创建者、日期、出版商及标识符等,更详细的还可提供摘要、关键词等元数据。描述的细粒度取决于学科、机构库及数据存储库的具体规定及要求,不过最好能按照 DataCite 元数据方案描述数据集[332]。该方案是一个跨学科的发现元数据标准,有助于数据集的描述与发现。如果是特定学科领域的机构库及数据存储库,最好能按照学科领域标准元数据著录数据集。DCC 提供了多种学科的元数据标准,各机构库及数据存储库可以此为参考,要求数据创建者提供最基本的元数据[33]。其次,机构库及数据存储库还应提供展示型元数据,帮助用户理解及重用数据,这其中最重要的是能帮助用户复制及验证研究成果。如在实验领域,提供的元数据应能支持用户利用该数据开展实验可得出相同的结论,在观测领域,提供的元数据应能支持用户从原始数据中得出相同的结论或利用新数据集开展研究,进而与原始结论进行对比。笔者在调查 7 个数据存储库时,发现它们大都提供了最基本的发现型元数据,但是大多还未提供详细的展示型元数据,较少提供关于数据收集过程以及如何应用的详细说明,有的数据存储库仅提供数据集所在项目的介绍,如 ICPSR、Research Data Australia。因此,机构库及数据存储库在未来需要加强展示型元数据的提供,帮助用户理解及重用数据,进而增加数据的使用量。最后,机构库及数据存储库应提供数据引用的标准格式,为数据引用的计量提供便利,建立基于使用及引用的综合数据计量。

6.4 出版商(数据期刊)角度

作为出版商,如何支持开放数据并激励研究人员像对待文章一样对待其研究数据[34]?首先,应制定政策鼓励科研人员存储其研究数据至一个能给数据集分配永久、可引用的标识符的稳定存储库。其次,指导研究人员在其参考文献列表中正确引用自己的研究数据或其他相关的数据;第三,在 CrossRef 中索引数据引用。CrossRef 在推动学术文献的关联、引用及检索上发挥了重要作用,其除了保证学术文献能通过 DOI 永久存取外,还提供基础设施使出版商在出版物出版时能够存储出版物与其相关资源的元数据,以保证出版物能够有效的查找、引用、链接与评估。随着研究数据的出现,出版物与数据的关联(如数据引用)成为该项服务的新内容之一。CrossRef 建议出版商在提交内容注册元数据时以参考文献或关系类型存储数据引用信息,这样期刊及出版商之间的数据引用将得以整合并通过单一门户供学术社区检索与使用。为此,CrossRef 制定了相关操作指南[35],规定出版商可在参考文献或关系类型两个地方实现与数据的关联。其中,以参考文献方式关联数据是指出版商将数据引用信息添加至

每个出版物的参考文献列表中;以关系类型关联数据是指出版商将数据的链接插入至元数据存储的"relationship"字段区,该字段能实现出版物与研究数据及其他相关资源的关联。

6.5 科研资助机构和科研机构角度

科研资助机构和科研机构是推动数据共享与开放的顶层组织,也是推动数据计量的关键者。科研资助机构和科研机构主要是在政策层面推动数据共享与开放。目前科研资助机构和科研机构的数据共享政策较少提及以数据计量的激励方式鼓励科研人员数据共享,而多项研究也指出,数据发布与共享激励政策的缺乏被认为是建立数据计量文化的一个障碍。为此,科研机构和科研资助机构需要为科研人员提供包括数据计量的激励体系,开发一套数据计量指标用来衡量共享数据的贡献,将其作为学术研究环境的一部分,在聘用、任职和升职决定中考虑数据共享活动。科研机构可有两种选择:一种是仅针对数据的独立激励体系,包括数据使用、引用及替代计量等;第二种选择是将数据计量和现有的科研激励体系结合起来,作为现有评估系统的补充。具体采用何种方式需要科研资助机构和科研机构结合自身情况在实践中逐步验证。

6.6 图书馆角度

图书馆在以文献为主体的传统科研评价体系中具有重要角色与作用,是科研评价数据,如引用数据的主要提供者,也是文献计量方法的主要应用者。随着研究数据越来越成为学术产出的重要组成部分,图书馆仍可利用其熟悉文献计量分析方法及科研评价指标与数据库的优势提供数据计量服务:①向科研人员宣传数据引用,指导科研人员规范数据引用行为;②利用数据引用索引及替代计量工具为科研人员及科研机构提供相应的数据引用及替代计量数据;③已建设数据存储库的图书馆可尝试按照《研究数据的使用统计实施规范》提供研究数据的使用统计分析报告;④推动机构层面的数据计量,与机构合作开发并测试合适的计量方法;⑤开发数据存储库的发现工具,推动数据的查找和利用。

7 结语

《研究数据的使用统计实施规范》通过对研究数据的使用统计的产生及传递进行标准化规范,使得数据存储库及平台提供商能以一种标准化的格式提供使用报告,为数据计量与影响力评估提供了一种新指标与新视角。被各参与方接受与使用是检验一个标准是否成功的重要标志。除了需要标准组织努力宣传推广外,数据使用计量更多依赖于各利益相关者在数据产生、管理、传播与利用等环节建立较为完善的配套机制。相信在各利益相关者的合作下,随着数据使用、引用、替代计量等指标体系的完善,科学界将建立完整的研究数据评价标准,推动研究数据的共享与利用。

参考文献:

[1] KRATZ J E, STRASSER C. Making data count. [EB/OL]. [2018－12－13]. https://doi.org/10.1038/sdata.2015.39.

[2] FENNER M, LOWENBERG D, JONE M, et al. Code of practice for research data usage metrics, release 1[EB/OL]. [2018－12－13]. https://www.projectcounter.org/code-of-prac tice-rd-sections/foreword/.

[3] NISO. Outputs of the NISO alternative assessment metrics project (NISO RP－25－2016) [EB/OL]. [2018－12－29]. http://www.niso.org/apps/group_public/download.php/17091/NISO RP-25-2016 Outputs of the NISO Alternative Assessment Project.pdf.

[4] FEAR K M. Measuring and anticipating the impact of data reuse [EB/OL]. [2018－12－15]. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/102481/kfear_1.pdf? sequence=1.

[5] BOLLEN J, DE SOMPEL H V, RODRIGUEZ M A, et al. Towards usage-based impact metrics: first results from the MESUR project[C] // Proceedings of the 8th ACM/IEEE-CS joint conference on digital libraries. New York: ACM, 2008:231－240.

[6] FORCE11. Joint declaration of data citation principles[EB/OL]. [2018－12－29]. https://www.force11.org/datacitationprinciples.

[7] The scholix initiative[EB/OL]. [2018－12－29]. http://www.scholix.org/.

[8] NISO. Alternative outputs in scholarly communications——data metrics(NISO RP-25-201x-2A) [EB/OL]. [2018－12－29]. https://groups.niso.org/apps/group_public/download.php/16553/NISO%20RP-25-201x-2A%20Alternative%20Outputs%20in%20Scholarly%20Communications--Data%20Metrics.pdf.

[9] Partnerships: PLOS, CDL, and DataONE launch pilot project to develop data-level metrics[EB/OL]. [2018－12－29]. http://www.infodocket.com/2014/10/07/partnerships-plos-cdl-and-dataone-launch-pilot-project-to-develop-data-level-metrics/.

[10] FEAR K M. The impact of data reuse: a pilot study of five measures[EB/OL]. [2018－12－30]. https://www.slideshare.net/asist_org/kfear-rdap.

[11] COSTAS R, MEIJER I, ZAHEDI Z, et al. The value of research data: metrics for datasets from a cultural and technical point of view [EB/OL]. [2018－12－30]. http://www.knowledge-exchange.info/event/value-reseach-data-metrics.

[12] BALL A, DUKE M. DCC how-to-guides: how to track the impact of research data with metrics[EB/OL]. [2018－12－30]. http://www.dcc.ac.uk/resources/how-guides.

[13] CASRAI. Research dataset-level metrics [EB/OL]. [2018－12－

30]. http://ref. casrai. org/Research_Dataset-Level_Metrics.
[14] 顾立平. 数据级别计量——概念辨析与实践进展[J]. 中国图书馆学报,2015,41(2): 56-71.
[15] 王毅萍,马建玲. 国外科学数据影响力研究进展[J]. 图书情报工作,2017,61(7):118-126.
[16] 孟阳,屈宝强. 数据计量与文献计量之间的对比研究[J]. 情报理论与实践,2017,40(11):139-144,138.
[17] 丁楠,黎娇,李文雨泽,等. 基于引用的科学数据评价研究[J]. 图书与情报,2014(5): 95-99.
[18] 邢红梅,吕先竞,刘文君,等. 基于DCI的社会学数据影响力分析[J]. 图书馆理论与实践,2016(2):43-46.
[19] INGWERSEN P, CHAVAN V. Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. BMC bioinformatics, 2011,12(Suppl 15):S3.
[20] 丁培. 科学数据使用统计应用及关键流程研究[J]. 现代情报,2017,37(7):116-122.
[21] 李洪. 新版COUNTER的特征及未来发展[J]. 中国图书馆学报,2012,38(6):29-37.
[22] DEKKERS A, ISAAC A. Data Catalog Vocabulary (DCAT) 1.1[EB/OL]. [2018-12-13]. https://w3c. github. io/dxwg/dcat/.
[23] COUNTER Code of Practice Release 5[EB/OL]. [2018-12-13]. https://www. projectcounter. org/code-of-practice-five-sections/foreword/.
[24] ANSI/NISO Z39.93-2014 SUSHI Protocol [EB/OL]. [2018-12-13]. http://www. niso. org/apps/group_public/download. php/14217/Z39-93-2014_SUSHI-1_7. pdf.
[25] BURTON A, FENNER M, HAAK W, et al. Scholix metadata schema for exchange of scholarly communication links[EB/OL]. [2018-12-13]. https://doi. org/10. 5281/zenodo. 1120 265.
[26] How to make your data count[EB/OL]. [2018-12-25]. https://makedatacount. files. wordpress. com/2018/07/how-to-make-your-data-count-webinar-july-2018. pdf.
[27] Dash how-to-guideline[EB/OL]. [2018-12-24]. https://github. com/CDLUC3/Make-Data-Count/blob/master/getting-started. md.
[28] DataONE implements new usage and citation metrics to make your data count[EB/OL]. [2018-12-25]. https://makedatacount. org/2018/10/24/dataone-implements-new-usage-and-citation-metrics-to-make-your-data-count/.
[29] Zenodo usage statistics launched [EB/OL]. [2018-12-27]. http://blog. zenodo. org/2018/07/ 18/2018-07-18-usage-statistics/.
[30] California Digital Library. Implementing the COUNTER code of practice for research data in repositories[EB/OL]. [2018-12-21]. https://github. com/CDLUC3/Make-Data-Count/blob/master/getting-started. md.
[31] WHYTE A. Where to keep your data: key considerations (checklist). [EB/OL]. [2018-12-21]. http://www. dcc. ac. uk/resources/how-guides/.
[32] DataCite Metadata Working Group. DataCite metadata schema for the publication and citation of research data [EB/OL]. [2018-12-21]. http://doi. org/10. 5438/0010. [33] DCC disciplinary metadata standards[EB/OL]. [2018-12-21]. http://www. dcc. ac. uk/resources/metadata-standards.
[34] Publishers: make your data citations count! [EB/OL]. [2018-12-25]. https://makedatacount. org/2018/05/29/publishers-make-your-data-citations-count/.
[35] Crossref data & software citation deposit guide for publishers [EB/OL]. [2018-12-25]. https://support. crossref. org/hc/en-us/articles/215787303-Crossref-Data-Software-Citation-Deposit-Guide-for-Publishers.

作者贡献说明:

林伟明:负责论文撰写;

叶兰:负责论文构思及修改指导。

Research on the Code of Practice for Research Data Usage Metrics and Its Implementation

Lin Weiming Ye Lan
Shenzhen University Library, Shenzhen 518060

Abstract: [**Purpose/significance**] This paper introduces the Code of Practice for Research Data Usage Metrics developed by Make Data Count and COUNTER, to provide a new metric for evaluating research data impacts from a new perspective. [**Method/process**] Through the analysis of the code, the background, purpose, scope, definitions of data elements and other terms, and core contents were introduced. Then, through the case analysis, the application of the specification by Dash, DataONE, Zenodo and seven other data repositories were surveyed. [**Result/conclusion**] The usage of research data has its own unique features. The introduction of the Code can complent the work of data citation and altmetrics for data, and further measure the research impact from historic perspective. At present, there are not many data repositories that follow this standard. In order to promote the application of research data usage metrics, standards organizations, researchers, repositories, publishers, funders, research institutions and libraries should cooperate in the production, management, dissemination and utilization of research data.

Keywords: research data usage metrics usage statistics usage data data-level metrics